Chapter 6

Monte Carlo Methods for Inferential Statistics

6.1 Introduction

Methods in inferential statistics are used to draw conclusions about a population and to measure the reliability of these conclusions using information obtained from a random sample. Inferential statistics involves techniques such as estimating population parameters using point estimates, calculating confidence interval estimates for parameters, hypothesis testing, and modeling (e.g., regression and density estimation). To measure the reliability of the inferences that are made, the statistician must understand the distribution of any statistics that are used in the analysis. In situations where we use a well-understood statistic, such as the sample mean, this is easily done analytically. However, in many applications, we do not want to be limited to using such simple statistics or to making simplifying assumptions. The goal of this chapter is to explain how simulation or Monte Carlo methods can be used to make inferences when the traditional or analytical statistical methods fail.

According to Murdoch [2000], the term ***Monte Carlo*** originally referred to simulations that involved random walks and was first used by Jon von Neumann and S. M. Ulam in the 1940's. Today, the ***Monte Carlo method*** refers to any simulation that involves the use of random numbers. In the following sections, we show that Monte Carlo simulations (or experiments) are an easy and inexpensive way to understand the phenomena of interest [Gentle, 1998]. To conduct a simulation experiment, you need a model that represents your population or phenomena of interest and a way to generate random numbers (according to your model) using a computer. The data that are generated from your model can then be studied as if they were observations. As we will see, one can use statistics based on the simulated data (means, medians, modes, variance, skewness, etc.) to gain understanding about the population.

In Section 6.2, we give a short overview of methods used in classical inferential statistics, covering such topics as hypothesis testing, power, and confidence intervals. The reader who is familiar with these may skip this section. In Section 6.3, we discuss Monte Carlo simulation methods for hypothesis testing and for evaluating the performance of the tests. The bootstrap method for estimating the bias and variance of estimates is presented in Section 6.4. Finally, Sections 6.5 and 6.6 conclude the chapter with information about available MATLAB code and references on Monte Carlo simulation and the bootstrap.

6.2 Classical Inferential Statistics

In this section, we will cover two of the main methods in inferential statistics: hypothesis testing and calculating confidence intervals. With confidence intervals, we are interested in obtaining an interval of real numbers that we expect (with specified confidence) contains the true value of a population parameter. In hypothesis testing, our goal is to make a decision about not rejecting or rejecting some statement about the population based on data from a random sample. We give a brief summary of the concepts in classical inferential statistics, endeavoring to keep the theory to a minimum. There are many books available that contain more information on these topics. We recommend Casella and Berger [1990], Walpole and Myers [1985], Bickel and Doksum [1977], Lindgren [1993], Montgomery, Runger and Hubele [1998], and Mood, Graybill and Boes [1974].

Hypothesis Testing

In hypothesis testing, we start with a ***statistical hypothesis***, which is a conjecture about one or more populations. Some examples of these are:

- A transportation official in the Washington, D.C. area thinks that the mean travel time to work for northern Virginia residents has increased from the average time it took in 1995.
- A medical researcher would like to determine whether aspirin decreases the risk of heart attacks.
- A pharmaceutical company needs to decide whether a new vaccine is superior to the one currently in use.
- An engineer has to determine whether there is a difference in accuracy between two types of instruments.

We generally formulate our statistical hypotheses in two parts. The first is the ***null hypothesis*** represented by H_0, which denotes the hypothesis we would like to test. Usually, we are searching for departures from this statement. Using one of the examples given above, the engineer would have the null hypothesis that there is no difference in the accuracy between the two instruments.

There must be an ***alternative hypothesis*** such that we would decide in favor of one or the other, and this is denoted by H_1. If we reject H_0, then this leads to the acceptance of H_1. Returning to the engineering example, the alternative hypothesis might be that there is a difference in the instruments or that one is more accurate than the other. When we perform a statistical hypothesis test, we can never know with certainty what hypothesis is true. For ease of exposition, we will use the terms ***accept the null hypothesis*** and ***reject the null hypothesis*** for our decisions resulting from statistical hypothesis testing.

To clarify these ideas, let's look at the example of the transportation official who wants to determine whether the average travel time to work has increased from the time it took in 1995. The mean travel time to work for northern Virginia residents in 1995 was 45 minutes. Since he wants to determine whether the mean travel time has increased, the statistical hypotheses are given by:

$$H_0: \qquad \mu = 45 \text{ minutes}$$

$$H_1: \qquad \mu > 45 \text{ minutes.}$$

The logic behind statistical hypothesis testing is summarized below, with details and definitions given after.

STEPS OF HYPOTHESIS TESTING

1. Determine the null and alternative hypotheses, using mathematical expressions if applicable. Usually, this is an expression that involves a characteristic or descriptive measure of a population.
2. Take a random sample from the population of interest.
3. Calculate a statistic from the sample that provides information about the null hypothesis. We use this to make our decision.
4. If the value of the statistic is consistent with the null hypothesis, then do not reject H_0.
5. If the value of the statistic is not consistent with the null hypothesis, then reject H_0 and accept the alternative hypothesis.

The problem then becomes one of determining when a statistic is consistent with the null hypothesis. Recall from Chapter 3 that a statistic is itself a random variable and has a probability distribution associated with it. So, in order to decide whether or not an observed value of the statistic is consistent with the null hypothesis, we must know the distribution of the statistic when the null hypothesis is true. The statistic used in step 3 is called a ***test statistic***.

Let's return to the example of the travel time to work for northern Virginia residents. To perform the analysis, the transportation official takes a random sample of 100 residents in northern Virginia and measures the time it takes them to travel to work. He uses the sample mean to help determine whether there is sufficient evidence to reject the null hypothesis and conclude that the mean travel time has increased. The sample mean that he calculates is 47.2 minutes. This is slightly higher than the mean of 45 minutes for the null hypothesis. However, the sample mean is a random variable and has some variation associated with it. If the variance of the sample mean under the null hypothesis is large, then the observed value of $\bar{x} = 47.2$ minutes might not be inconsistent with H_0. This is explained further in Example 6.1.

Example 6.1

We continue with the transportation example. We need to determine whether or not the value of the statistic obtained from a random sample drawn from the population is consistent with the null hypothesis. Here we have a random sample comprised of $n = 100$ commute times. The sample mean of these observations is $\bar{x} = 47.2$ minutes. If the transportation official assumes that the travel times to work are normally distributed with $\sigma = 15$ minutes (one might know a reasonable value for σ based on previous experience with the population), then we know from Chapter 3 that $\bar{x}$ is approximately normally distributed with mean μ_X and standard deviation $\sigma_{\bar{X}} = \sigma_X / \sqrt{n}$. Standardizing the observed value of the sample mean, we have

$$z_o = \frac{\bar{x} - \mu_0}{\sigma_X / \sqrt{n}} = \frac{\bar{x} - \mu_0}{\sigma_{\bar{X}}} = \frac{47.2 - 45}{15 / \sqrt{100}} = \frac{2.2}{1.5} = 1.47 \,, \tag{6.1}$$

where z_o is the observed value of the test statistic, and μ_0 is the mean under the null hypothesis. Thus, we have that the value of $\bar{x} = 47.2$ minutes is 1.47 standard deviations away from the mean, if the null hypothesis is really true. (This is why we use μ_0 in Equation 6.1.) We know that approximately 95% of normally distributed random variables fall within two standard deviations either side of the mean. Thus, $\bar{x} = 47.2$ minutes is not inconsistent with the null hypothesis.

❑

In hypothesis testing, the rule that governs our decision might be of the form: *if the observed statistic is within some region, then we reject the null hypothesis*. The ***critical region*** is an interval for the test statistic over which we would reject H_0. This is sometimes called the ***rejection region***. The ***critical value*** is that value of the test statistic that divides the domain of the test statistic into a region where H_0 will be rejected and one where H_0 will be accepted. We need to know the distribution of the test statistic under the null hypothesis to find the critical value(s).

The critical region depends on the distribution of the statistic under the null hypothesis, the alternative hypothesis, and the amount of error we are willing to tolerate. Typically, the critical regions are areas in the tails of the distribution of the test statistic when H_0 is true. It could be in the lower tail, the upper tail or both tails, and which one is appropriate depends on the alternative hypothesis. For example:

- If a large value of the test statistic would provide evidence for the alternative hypothesis, then the critical region is in the upper tail of the distribution of the test statistic. This is sometimes referred to as an ***upper tail test***.
- If a small value of the test statistic provides evidence for the alternative hypothesis, then the critical region is in the lower tail of the distribution of the test statistic. This is sometimes referred to as a ***lower tail test***.
- If small or large values of the test statistic indicate evidence for the alternative hypothesis, then the critical region is in the lower and upper tails. This is sometimes referred to as a ***two-tail test***.

There are two types of errors that can occur when we make a decision in statistical hypothesis testing. The first is a ***Type I error***, which arises when we reject H_0 when it is really true. The other error is called ***Type II error***, and this happens when we fail to detect that H_0 is actually false. These errors are summarized in Table 6.1.

TABLE 6.1

Types of Error in Statistical Hypothesis Testing

Type of Error	**Description**	**Probability of Error**
Type I Error	Rejecting H_0 when it is true	α
Type II Error	Not rejecting H_0 when it is false	β

Recall that we are usually searching for significant evidence that the alternative hypothesis is valid, and we do not want to change from the status quo (i.e., reject H_0) unless there is sufficient evidence in the data to lead us in that direction. So, when setting up a hypothesis test we ensure that the probability of wrongly rejecting H_0 is controlled. The probability of making a Type I error is denoted by α and is sometimes called the ***significance level*** of the test. The α is set by the analyst, and it represents the maximum probability of Type I error that will be tolerated. Typical values of α are $\alpha = 0.01, 0.05, 0.10$. The critical value is found as the quantile (under the null hypothesis) that gives a significance level of α.

The specific procedure for conducting an hypothesis test using these ideas is given below. This is called the ***critical value approach***, because the decision is based on whether the value of the test statistic falls in the rejection region. We will discuss an alternative method later in this section. The concepts of hypothesis testing using the critical value approach are illustrated in Example 6.2.

PROCEDURE - HYPOTHESIS TESTING (CRITICAL VALUE APPROACH)

1. Determine the null and alternative hypotheses.
2. Find a test statistic T that will provide evidence that H_0 should be accepted or rejected (e.g, a large value of the test statistic indicates H_0 should be rejected).
3. Obtain a random sample from the population of interest and compute the observed value of the test statistic t_o using the sample.
4. Using the sampling distribution of the test statistic under the null hypothesis and the significance level, find the critical value(s). That is, find the t such that

   Upper Tail Test: $P_{H_0}(T \le t) = 1 - \alpha$

   Lower Tail Test: $P_{H_0}(T \le t) = \alpha$

   Two-Tail Test: $P_{H_0}(T \le t_1) = \alpha/2$ and $P_{H_0}(T \le t_2) = 1 - \alpha/2$,

   where $P_{H_0}(.)$ denotes the probability under the null hypothesis.

5. If the value of the test statistic t_o falls in the critical region, then reject the null hypothesis.

Example 6.2

Here, we illustrate the critical value approach to hypothesis testing using the transportation example. Our test statistic is given by

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{X}}},$$

and we observed a value of $z_o = 1.47$ based on the random sample of $n = 100$ commute times. We want to conduct the hypothesis test at a significance level given by $\alpha = 0.05$. Since our alternative hypothesis is that the commute times have increased, a large value of the test statistic provides evidence for H_1. We can find the critical value using the MATLAB Statistics Toolbox as follows:

```
cv = norminv(0.95,0,1);
```

This yields a critical value of 1.645. Thus, if $z_o \geq 1.645$, then we reject H_0. Since the observed value of the test statistic is less than the critical value, we do not reject H_0. The regions corresponding to this hypothesis test are illustrated in Figure 6.1.

❑

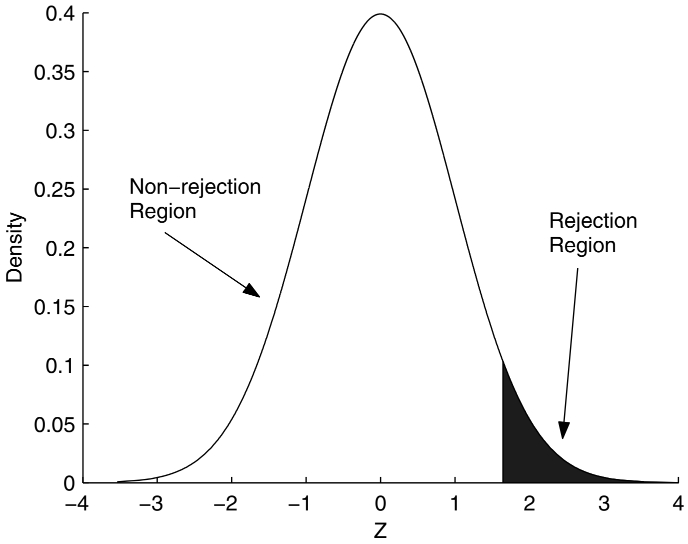

FIGURE 6.1
This shows the critical region (shaded region) for the hypothesis test of Examples 6.1 and 6.2. If the observed value of the test statistic falls in the shaded region, then we reject the null hypothesis. Note that this curve reflects the distribution for the test statistic under the null hypothesis.

The probability of making a Type II error is represented by β, and it depends on the sample size, the significance level of the test, and the alternative hypothesis. The last part is important to remember: *the probability that we will not detect a departure from the null hypothesis depends on the distribution of the test statistic under the alternative hypothesis*. Recall that the alternative hypothesis allows for many different possibilities, yielding many distributions under H_1. So, we must determine the Type II error for every alternative hypothesis of interest.

A more convenient measure of the performance of a hypothesis test is to determine the probability of not making a Type II error. This is called the ***power*** of a test. We can consider this to be the probability of rejecting H_0 when it is really false. Roughly speaking, one can think of the power as the ability of the hypothesis test to detect a false null hypothesis. The power is given by

$$\text{Power} = 1 - \beta . \tag{6.2}$$

As we see in Example 6.3, the power of the test to detect departures from the null hypothesis depends on the true value of μ .

Example 6.3

Returning to the transportation example, we illustrate the concepts of Type II error and power. It is important to keep in mind that these values depend on the true mean μ, so we have to calculate the Type II error for different values of μ. First we get a vector of values for μ:

```
% Get several values for the mean under the alternative
% hypothesis. Note that we are getting some values
% below the null hypothesis.
mualt = 40:60;
```

It is actually easier to understand the power when we look at a test statistic based on $\bar{x}$ rather than z_o. So, we convert the critical value to its corresponding $\bar{x}$ value:

```
% Note the critical value:
cv = 1.645;
% Note the standard deviation for x-bar:
sig = 1.5;
% It's easier to use the non-standardized version,
% so convert:
ct = cv*1.5 + 45;
```

We find the area under the curve to the left of the critical value (the non rejection region) for each of these values of the true mean. That would be the probability of not rejecting the null hypothesis.

```
% Get a vector of critical values that is
% the same size as mualt.
ctv = ct*ones(size(mualt));
% Now get the probabilities to the left of this value.
% These are the probabilities of the Type II error.
beta = normcdf(ctv,mualt,sig);
```

Note that the variable `beta` contains the probability of Type II error (the area to the left of the critical value `ctv` under a normal curve with mean `mualt` and standard deviation `sig`) for every μ .To get the power, simply subtract all of the values for `beta` from one.

```
% To get the power: 1-beta
```

```
pow = 1 - beta;
```

We plot the power against the true value of the population mean in Figure 6.2. Note that as $\mu > \mu_0$, the power (or the likelihood that we can detect the alternative hypothesis) increases.

```
plot(mualt,pow);
xlabel('True Mean \mu')
ylabel('Power')
axis([40 60 0 1.1])
```

We leave it as an exercise for the reader to plot the probability of making a Type II error.

❑

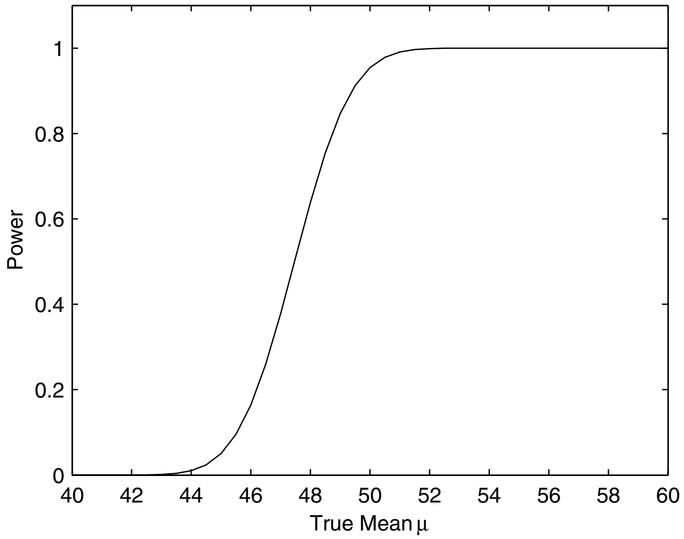

FIGURE 6.2
This shows the power (or probability of not making aType II error) as a function of the true value of the population mean μ. Note that as the true mean gets larger, then the likelihood of not making a Type II error increases.

There is an alternative approach to hypothesis testing, which uses a quantity called a *p*-value. A ***p-value*** is defined as the probability of observing a value of the test statistic as extreme as or more extreme than the one that is observed, when the null hypothesis H_0 is true. The word *extreme* refers to the direction of the alternative hypothesis. For example, if a small value of the test statistic (a lower tail test) indicates evidence for the alternative hypothesis, then the *p*-value is calculated as

$$p - \text{value} = P_{H_0}(T \le t_o),$$

where t_o is the observed value of the test statistic T, and $P_{H_0}(.)$ denotes the probability under the null hypothesis. The p-value is sometimes referred to as the ***observed significance level***.

In the p-value approach, a small value indicates evidence for the alternative hypothesis and would lead to rejection of H_0. Here small refers to a p-value that is less than or equal to α. The steps for performing hypothesis testing using the p-value approach are given below and are illustrated in Example 6.4.

PROCEDURE - HYPOTHESIS TESTING (P-VALUE APPROACH)

1. Determine the null and alternative hypotheses.
2. Find a test statistic T that will provide evidence about H_0.
3. Obtain a random sample from the population of interest and compute the value of the test statistic t_o from the sample.
4. Calculate the p-value:

   Lower Tail Test: $p - \text{value} = P_{H_0}(T \le t_o)$

   Upper Tail Test: $p - \text{value} = P_{H_0}(T \ge t_o)$

5. If the $p\text{-value} \le \alpha$, then reject the null hypothesis.

For a two-tail test, the p-value is determined similarly.

Example 6.4

In this example, we repeat the hypothesis test of Example 6.2 using the p-value approach. First we set some of the values we need:

```
mu = 45;
sig = 1.5;
xbar = 47.2;
% Get the observed value of test statistic.
zobs = (xbar - mu)/sig;
```

The p-value is the area under the curve greater than the value for **zobs**. We can find it using the following command:

```
pval = 1-normcdf(zobs,0,1);
```

We get a p-value of 0.071. If we are doing the hypothesis test at the 0.05 significance level, then we would not reject the null hypothesis. This is consistent with the results we had previously.

❑

Note that in each approach, knowledge of the distribution of T under the null hypothesis H_0 is needed. How to tackle situations where we do not know the distribution of our statistic is the focus of the rest of the chapter.

Confidence Intervals

In Chapter 3, we discussed several examples of estimators for population parameters such as the mean, the variance, moments, and others. We call these ***point estimates.*** It is unlikely that a point estimate obtained from a random sample will exactly equal the true value of the population parameter. Thus, it might be more useful to have an interval of numbers that we expect will contain the value of the parameter. This type of estimate is called an ***interval estimate***. An understanding of confidence intervals is needed for the bootstrap methods covered in Section 6.4.

Let θ represent a population parameter that we wish to estimate, and let T denote a statistic that we will use as a point estimate for θ. The observed value of the statistic is denoted as $\hat{\theta}$. An interval estimate for θ will be of the form

$$\hat{\theta}_{Lo} < \theta < \hat{\theta}_{Up}\,, \tag{6.3}$$

where $\hat{\theta}_{Lo}$ and $\hat{\theta}_{Up}$ depend on the observed value $\hat{\theta}$ and the distribution of the statistic T.

If we know the sampling distribution of T, then we are able to determine values for $\hat{\theta}_{Lo}$ and $\hat{\theta}_{Up}$ such that

$$P(\hat{\theta}_{Lo} < \theta < \hat{\theta}_{Up}) = 1 - \alpha\,, \tag{6.4}$$

where $0 < \alpha < 1$. Equation 6.4 indicates that we have a probability of $1 - \alpha$ that we will select a random sample that produces an interval that contains θ. This interval (Equation 6.3) is called a $(1 - \alpha) \cdot 100\%$ confidence interval. The philosophy underlying confidence intervals is the following. Suppose we repeatedly take samples of size n from the population and compute the random interval given by Equation 6.3. Then the relative frequency of the intervals that contain the parameter θ would approach $(1 - \alpha) \cdot 100\%$. It should be noted that one-sided confidence intervals can be defined similarly [Mood, Graybill and Boes, 1974].

To illustrate these concepts, we use Equation 6.4 to get a confidence interval for the population mean μ. Recall from Chapter 3 that we know the distribution for $\overline{X}$. We define $z^{(\alpha/2)}$ as the z value that has an area under the standard normal curve of size $\alpha/2$ to the left of it. In other words, we use $z^{(\alpha/2)}$ to denote that value such that

$$P(Z < z^{(\alpha/2)}) = \alpha/2 .$$

Thus, the area between $z^{(\alpha/2)}$ and $z^{(1-\alpha/2)}$ is $1-\alpha$. This is shown in Figure 6.3.

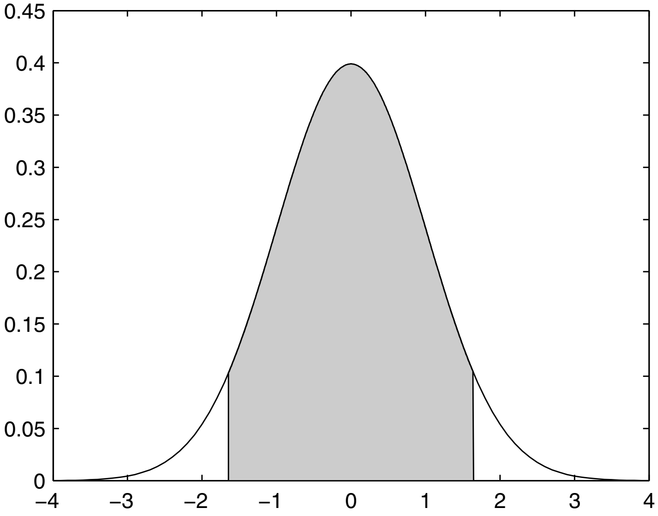

FIGURE 6.3
The left vertical line corresponds to $z^{(\alpha/2)}$, and the right vertical line is at $z^{(1-\alpha/2)}$. So, the non-shaded areas in the tails each have an area of $\alpha/2$, and the shaded area in the middle is $1-\alpha$.

We can see from this that the shaded area has probability $1-\alpha$, and

$$P(z^{(\alpha/2)} < Z < z^{(1-\alpha/2)}) = 1-\alpha , \tag{6.5}$$

where

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} . \tag{6.6}$$

If we substitute this into Equation 6.5, then we have

$$P\left(z^{(\alpha/2)} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z^{(1-\alpha/2)}\right) = 1 - \alpha . \tag{6.7}$$

Rearranging the inequalities in Equation 6.7, we obtain

$$P\left(\overline{X} - z^{(1-\alpha/2)}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - z^{(\alpha/2)}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha . \tag{6.8}$$

Comparing Equations 6.8 and 6.4, we see that

$$\hat{\theta}_{Lo} = \overline{X} - z^{(1-\alpha/2)}\frac{\sigma}{\sqrt{n}} \qquad \hat{\theta}_{Up} = \overline{X} - z^{(\alpha/2)}\frac{\sigma}{\sqrt{n}} .$$

Example 6.5

We provide an example of finding a 95% confidence interval, using the transportation application of before. Recall that $n = 100$, $\bar{x} = 47.2$ minutes, and the standard deviation of the travel time to work is $\sigma = 15$ minutes. Since we want a 95% confidence interval, $\alpha = 0.05$.

```
mu = 45;
sig = 15;
n = 100;
alpha = 0.05;
xbar = 47.2;
```

We can get the endpoints for a 95% confidence interval as follows:

```
% Get the 95% confidence interval.
% Get the value for z_alpha/2.
zlo = norminv(1-alpha/2,0,1);
zhi = norminv(alpha/2,0,1);
thetalo = xbar - zlo*sig/sqrt(n);
thetaup = xbar - zhi*sig/sqrt(n);
```

We get a value of $\hat{\theta}_{Lo} = 44.26$ and $\hat{\theta}_{Up} = 50.14$.
❑

We return to confidence intervals in Section 6.4 and Chapter 7, where we discuss bootstrap methods for obtaining them. First, however, we look at Monte Carlo methods for hypothesis testing.

6.3 Monte Carlo Methods for Inferential Statistics

The sampling distribution is known for many statistics. However, these are typically derived using assumptions about the underlying population under study or for large sample sizes. In many cases, we do not know the sampling distribution for the statistic, or we cannot be sure that the assumptions are satisfied. We can address these cases using Monte Carlo simulation methods, which is the topic of this section. Some of the uses of Monte Carlo simulation for inferential statistics are the following:

- Performing inference when the distribution of the test statistic is not known analytically,
- Assessing the performance of inferential methods when parametric assumptions are violated,
- Testing the null and alternative hypotheses under various conditions,
- Evaluating the performance (e.g., power) of inferential methods,
- Comparing the quality of estimators.

In this section, we cover situations in inferential statistics where we do know something about the distribution of the population our sample came from or we are willing to make assumptions about the distribution. In Section 6.4, we discuss bootstrap methods that can be used when no assumptions are made about the underlying distribution of the population.

Basic Monte Carlo Procedure

The fundamental idea behind ***Monte Carlo simulation*** for inferential statistics is that insights regarding the characteristics of a statistic can be gained by repeatedly drawing random samples from the same population of interest and observing the behavior of the statistic over the samples. In other words, we estimate the distribution of the statistic by randomly sampling from the population and recording the value of the statistic for each sample. The observed values of the statistic for these samples are used to estimate the distribution.

The first step is to decide on a pseudo-population that the analyst assumes represents the real population in all relevant aspects. We use the word *pseudo* here to emphasize the fact that we obtain our samples using a computer and pseudo random numbers. For example, we might assume that the underlying population is exponentially distributed if the random variable represents the time before a part fails, or we could assume the random variable comes from a normal distribution if we are measuring IQ scores. The pseudo-population must be something we can sample from using the computer. In this text, we consider this type of Monte Carlo simulation to be a parametric technique, because we sample from a known or assumed distribution.

The basic Monte Carlo procedure is outlined here. Later, we provide procedures illustrating some specific uses of Monte Carlo simulation as applied to statistical hypothesis testing.

PROCEDURE - BASIC MONTE CARLO SIMULATION

1. Determine the pseudo-population or model that represents the true population of interest.
2. Use a sampling procedure to sample from the pseudo-population.
3. Calculate a value for the statistic of interest and store it.
4. Repeat steps 2 and 3 for M trials.
5. Use the M values found in step 4 to study the distribution of the statistic.

It is important to keep in mind, that when sampling from the pseudo-population, the analyst should ensure that all relevant characteristics reflect the statistical situation. For example, the same sample size and sampling strategy should be used when trying to understand the performance of a statistic. This means that the distribution for the statistic obtained via Monte Carlo simulation is valid only for the conditions of the sampling procedure and the assumptions about the pseudo-population.

Note that in the last step of the Monte Carlo simulation procedure, the analyst can use the estimated distribution of the statistic to study characteristics of interest. For example, one could use this information to estimate the skewness, bias, standard deviation, kurtosis and many other characteristics.

Monte Carlo Hypothesis Testing

Recall that in statistical hypothesis testing, we have a test statistic that provides evidence that the null hypothesis should be rejected or not. Once we observe the value of the test statistic, we decide whether or not that particular value is consistent with the null hypothesis. To make that decision, we must know the distribution of the statistic when the null hypothesis is true. Estimating the distribution of the test statistic under the null hypothesis is one of the goals of Monte Carlo hypothesis testing. We discuss and illustrate the Monte Carlo method as applied to the critical value and p-value approaches to hypothesis testing.

Recall that in the critical value approach to hypothesis testing, we are given a significance level α. We then use this significance level to find the appropriate critical region in the distribution of the test statistic when the null hypothesis is true. Using the Monte Carlo method, we determine the critical value using the estimated distribution of the test statistic. The basic procedure is to randomly sample many times from the pseudo-population representing the null hypothesis, calculate the value of the test statistic at each trial, and use these values to estimate the distribution of the test statistic.

PROCEDURE - MONTE CARLO HYPOTHESIS TESTING (CRITICAL VALUE)

1. Using an available random sample of size n from the population of interest, calculate the observed value of the test statistic, t_o.
2. Decide on a pseudo-population that reflects the characteristics of the true population under the null hypothesis.
3. Obtain a random sample of size n from the pseudo-population.
4. Calculate the value of the test statistic using the random sample in step 3 and record it.
5. Repeat steps 3 and 4 for M trials. We now have values $t_1, \ldots, t_M$, that serve as an estimate of the distribution of the test statistic, T, when the null hypothesis is true.
6. Obtain the critical value for the given significance level α:

   Lower Tail Test: get the α-th sample quantile, $\hat{q}_\alpha$, from the $t_1, \ldots, t_M$.
   Upper Tail Test: get the $(1-\alpha)$-th sample quantile, $\hat{q}_{1-\alpha}$, from the $t_1, \ldots, t_M$.
   Two-Tail Test: get the sample quantiles $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ from the $t_1, \ldots, t_M$.

7. If t_o falls in the critical region, then reject the null hypothesis.

The critical values in step 6 can be obtained using the estimate of a sample quantile that we discussed in Chapter 3. The function **`csquantiles`** from the Computational Statistics Toolbox is also available to find these values.

In the examples given below, we apply the Monte Carlo method to a familiar hypothesis testing situation where we are testing an hypothesis about the population mean. As we saw earlier, we can use analytical approaches for this type of test. We use this simple application in the hope that the reader will better understand the ideas of Monte Carlo hypothesis testing and then easily apply them to more complicated problems.

Example 6.6

This toy example illustrates the concepts of Monte Carlo hypothesis testing. The **`mcdata`** data set contains 25 observations. We are interested in using these data to test the following null and alternative hypotheses:

$$\begin{aligned} H_0&: \quad \mu = 454 \\ H_1&: \quad \mu < 454. \end{aligned}$$

We will perform our hypothesis test using simulation to get the critical values. We decide to use the following as our test statistic

$$z = \frac{\bar{x} - 454}{\sigma/\sqrt{n}}.$$

First, we take care of some preliminaries.

```
% Load up the data.
load mcdata
n = length(mcdata);
% Population sigma is known.
sigma = 7.8;
sigxbar = sigma/sqrt(n);
% Get the observed value of the test statistic.
Tobs = (mean(mcdata)-454)/sigxbar;
```

The observed value of the test statistic is $t_o = -2.56$. The next step is to decide on a model for the population that generated our data. We suspect that the normal distribution with $\sigma = 7.8$ is a good model, and we check this assumption using a normal probability plot. The resulting plot in Figure 6.4 shows that we can use the normal distribution as the pseudo-population.

```
% This command generates the normal probability plot.
% It is a function in the MATLAB Statistics Toolbox.
normplot(mcdata)
```

We are now ready to implement the Monte Carlo simulation. We use 1000 trials in this example. At each trial, we randomly sample from the distribution of the test statistic under the null hypothesis (the normal distribution with $\mu = 454$ and $\sigma = 7.8$) and record the value of the test statistic.

```
M = 1000;% Number of Monte Carlo trials
% Storage for test statistics from the MC trials.
Tm = zeros(1,M);
% Start the simulation.
for i = 1:M
   % Generate a random sample under H_0
   % where n is the sample size.
   xs = sigma*randn(1,n) + 454;
   Tm(i) = (mean(xs) - 454)/sigxbar;
end
```

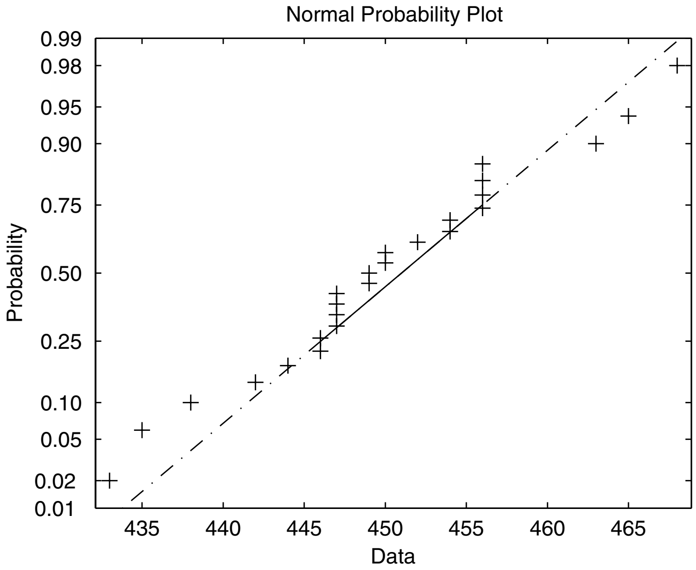

FIGURE 6.4
This normal probability plot for the **mcdata** data shows that assuming a normal distribution for the data is reasonable.

Now that we have the estimated distribution of the test statistic contained in the variable **Tm**, we can use that to estimate the critical value for a lower tail test.

```
% Get the critical value for alpha.
% This is a lower-tail test, so it is the
% alpha quantile.
alpha = 0.05;
cv = csquantiles(Tm,alpha);
```

We get an estimated critical value of -1.75. Since the observed value of our test statistic is $t_o = -2.56$, which is less than the estimated critical value, we reject H_0.
❑

The procedure for Monte Carlo hypothesis testing using the p-value approach is similar. Instead of finding the critical value from the simulated distribution of the test statistic, we use it to estimate the p-value.

PROCEDURE - MONTE CARLO HYPOTHESIS TESTING (P-VALUE)

1. For a random sample of size n to be used in a statistical hypothesis test, calculate the observed value of the test statistic, t_o.
2. Decide on a pseudo-population that reflects the characteristics of the population under the null hypothesis.
3. Obtain a random sample of size n from the pseudo-population.
4. Calculate the value of the test statistic using the random sample in step 3 and record it as t_i.
5. Repeat steps 3 and 4 for M trials. We now have values $t_1, \ldots, t_M$, that serve as an estimate of the distribution of the test statistic, T, when the null hypothesis is true.
6. Estimate the p-value using the distribution found in step 5, using the following.

   Lower Tail Test:

$$\hat{p}\text{-value} = \frac{\#(t_i \le t_o)}{M}; \qquad i = 1, \ldots, M$$

   Upper Tail Test:

$$\hat{p}\text{-value} = \frac{\#(t_i \ge t_o)}{M}; \qquad i = 1, \ldots, M$$

7. If $\hat{p}\text{-value} \le \alpha$, then reject the null hypothesis.

Example 6.7

We return to the situation in Example 6.6 and apply Monte Carlo simulation to the p-value approach to hypothesis testing. Just to change things a bit, we use the sample mean as our test statistic.

```
% Let's change the test statistic to xbar.
Tobs = mean(mcdata);
% Number of Monte Carlo trials.
M = 1000;
% Start the simulation.
Tm = zeros(1,M);
for i = 1:M
   % Generate a random sample under H_0.
   xs = sigma*randn(1,n) + 454;
   Tm(i) = mean(xs);
end
```

We find the estimated p-value by counting the number of observations in **Tm** that are below the value of the observed value of the test statistic and dividing by M.

```
% Get the p-value. This is a lower tail test.
% Find all of the values from the simulation that are
% below the observed value of the test statistic.
ind = find(Tm <= Tobs);
pvalhat = length(ind)/M;
```

We have an estimated p-value given by 0.007. If the significance level of our test is $\alpha = 0.05$, then we would reject the null hypothesis.
❑

Monte Carlo Assessment of Hypothesis Testing

Monte Carlo simulation can be used to evaluate the performance of an inference model or hypothesis test in terms of the Type I error and the Type II error. For some statistics, such as the sample mean, these errors can be determined analytically. However, what if we have an inference test where the assumptions of the standard methods might be violated or the analytical methods cannot be applied? For instance, suppose we choose the critical value by using a normal approximation (when our test statistic is *not* normally distributed), and we need to assess the results of doing that? In these situations, we can use Monte Carlo simulation to estimate the Type I and the Type II error.

We first outline the procedure for estimating the Type I error. Because the Type I error occurs when we reject the null hypothesis test when it is true, we must sample from the pseudo-population that represents H_0.

PROCEDURE - MONTE CARLO ASSESSMENT OF TYPE I ERROR

1. Determine the pseudo-population when the null hypothesis is *true*.
2. Generate a random sample of size n from this pseudo-population.
3. Perform the hypothesis test using the critical value.
4. Determine whether a Type I error has been committed. In other words, was the null hypothesis rejected? We know that it should not be rejected because we are sampling from the distribution according to the null hypothesis. Record the result for this trial as,

$$I_i = \begin{cases} 1; & \text{Type I error is made} \\ 0; & \text{Type I error is not made.} \end{cases}$$

5. Repeat steps 2 through 4 for M trials.

6. The probability of making a Type I error is

$$\hat{\alpha} = \frac{1}{M}\sum_{i=1}^{M} I_i . \tag{6.9}$$

Note that in step 6, this is the same as calculating the proportion of times the null hypothesis is falsely rejected out of M trials. This provides an estimate of the significance level of the test for a given critical value.

The procedure is similar for estimating the Type II error of a hypothesis test. However, this error is determined by sampling from the distribution when the null hypothesis is false. There are many possibilities for the Type II error, and the analyst should investigate the Type II error for those alternative hypotheses that are of interest.

PROCEDURE - MONTE CARLO ASSESSMENT OF TYPE II ERROR

1. Determine a pseudo-population of interest where the null hypothesis is *false*.
2. Generate a random sample of size n from this pseudo-population.
3. Perform the hypothesis test using the significance level α and corresponding critical value.
4. Note whether a Type II error has been committed; i.e., was the null hypothesis *not* rejected? Record the result for this trial as,

$$I_i = \begin{cases} 1; & \text{Type II error is made} \\ 0; & \text{Type II error is not made.} \end{cases}$$

5. Repeat steps 2 through 4 for M trials.
6. The probability of making a Type II error is

$$\hat{\beta} = \frac{1}{M}\sum_{i=1}^{M} I_i . \tag{6.10}$$

The Type II error rate is estimated using the proportion of times the null hypothesis is not rejected (when it should be) out of M trials.

Example 6.8

For the hypothesis test in Example 6.6, we had a critical value (from theory) of -1.645. We can estimate the significance level of the test using the following steps:

```
M = 1000;
alpha = 0.05;
% Get the critical value, using z as test statistic.
cv = norminv(alpha,0,1);
% Start the simulation.
Im = 0;
for i = 1:M
 % Generate a random sample under H_0.
 xs = sigma*randn(1,n) + 454;
 Tm = (mean(xs)-454)/sigxbar;
 if Tm <= cv   % then reject H_0
   Im = Im +1;
 end
end
alphahat = Im/M;
```

A critical value of -1.645 in this situation corresponds to a desired probability of Type I error of 0.05. From this simulation, we get an estimated value of 0.045, which is very close to the theoretical value. We now check the Type II error in this test. Note that we now have to sample from the alternative hypotheses of interest.

```
% Now check the probability of Type II error.
% Get some alternative hypotheses:
mualt = 445:458;
betahat = zeros(size(mualt));
for j = 1:length(mualt)
   Im = 0;
   % Get the true mean.
   mu = mualt(j);
   for i = 1:M
      % Generate a sample from H_1.
      xs = sigma*randn(1,n) + mu;
      Tm = (mean(xs)-454)/sigxbar;
      if Tm > cv   % Then did not reject H_0.
         Im = Im +1;
      end
   end
   betahat(j) = Im/M;
end
% Get the estimated power.
powhat = 1-betahat;
```

We plot the estimated power as a function of μ in Figure 6.5. As expected, as the true value for μ gets closer to 454 (the mean under the null hypothesis), the power of the test decreases.

❑

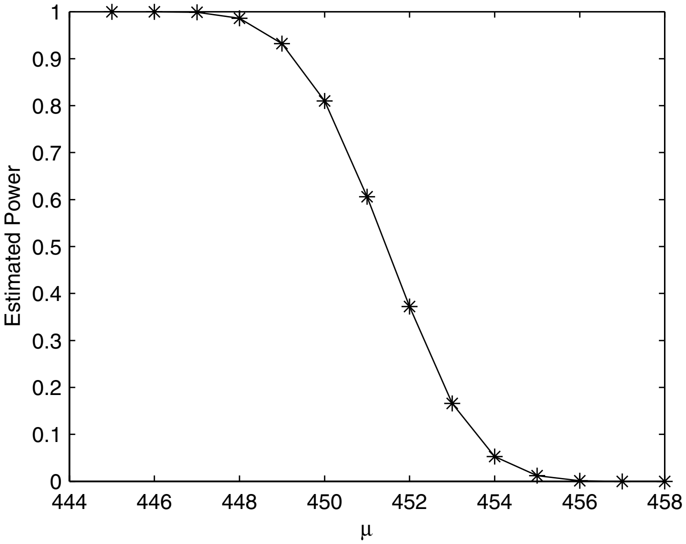

FIGURE 6.5
Here is the curve for the estimated power corresponding to the hypothesis test of Example 6.8.

An important point to keep in mind about the Monte Carlo simulations discussed in this section is that the experiment is applicable only for the situation that has been simulated. For example, when we assess the Type II error in Example 6.8, it is appropriate only for those alternative hypotheses, sample size and critical value. What would be the probability of Type II error, if some other departure from the null hypothesis is used in the simulation? In other cases, we might need to know whether the distribution of the statistic changes with sample size or skewness in the population or some other characteristic of interest. These variations are easily investigated using multiple Monte Carlo experiments.

One quantity that the researcher must determine is the number of trials that are needed in Monte Carlo simulations. This often depends on the computing assets that are available. If time and computer resources are not an issue, then M should be made as large as possible. Hope [1968] showed that results from a Monte Carlo simulation are unbiased for any M, under the assumption that the programming is correct.

Mooney [1997] states that there is no general theory that governs the number of trials in Monte Carlo simulation. However, he recommends the following general guidelines. The researcher should first use a small number of trials and ensure that the program is working properly. Once the code has been checked, the simulation or experiments can be run for very large M.

Most simulations would have $M > 1000$, but M between 10,000 and 25,000 is not uncommon. One important guideline for determining the number of trials, is the purpose of the simulation. If the tail of the distribution is of interest (e.g., estimating Type I error, getting p-values, etc.), then more trials are needed to ensure that there will be a good estimate of that area.

6.4 Bootstrap Methods

The treatment of the bootstrap methods described here comes from Efron and Tibshirani [1993]. The interested reader is referred to that text for more information on the underlying theory behind the bootstrap. There does not seem to be a consistent terminology in the literature for what techniques are considered bootstrap methods. Some refer to the resampling techniques of the previous section as bootstrap methods. Here, we use ***bootstrap*** to refer to Monte Carlo simulations that treat the original sample as the pseudo-population or as an estimate of the population. Thus, in the steps where we randomly sample from the pseudo-population, we now resample from the original sample.

In this section, we discuss the general bootstrap methodology, followed by some applications of the bootstrap. These include bootstrap estimates of the standard error, bootstrap estimates of bias, and bootstrap confidence intervals.

General Bootstrap Methodology

The bootstrap is a method of Monte Carlo simulation where no parametric assumptions are made about the underlying population that generated the random sample. Instead, we use the sample as an estimate of the population. This estimate is called the empirical distribution $\hat{F}$ where each x_i has probability mass $1/n$. Thus, each x_i has the same likelihood of being selected in a new sample taken from $\hat{F}$.

When we use $\hat{F}$ as our pseudo-population, then we resample *with replacement* from the original sample $\mathbf{x} = (x_1, \ldots, x_n)$. We denote the new sample obtained in this manner by $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$. Since we are sampling with replacement from the original sample, there is a possibility that some points x_i will appear more than once in $\mathbf{x}^*$ or maybe not at all. We are looking at the univariate situation, but the bootstrap concepts can also be applied in the d-dimensional case.

A small example serves to illustrate these ideas. Let's say that our random sample consists of the four numbers $\mathbf{x} = (5, 8, 3, 2)$. The following are possible samples $\mathbf{x}^*$, when we sample with replacement from $\mathbf{x}$:

$$\mathbf{x}^{*1} = (x_4, x_4, x_2, x_1) = (2, 2, 8, 5)$$

$$\mathbf{x}^{*2} = (x_4, x_2, x_3, x_4) = (2, 8, 3, 2).$$

We use the notation $\mathbf{x}^{*b}$, $b = 1, \ldots, B$ for the b-th bootstrap data set.

In many situations, the analyst is interested in estimating some parameter θ by calculating a statistic from the random sample. We denote this estimate by

$$\hat{\theta} = T = t(x_1, \ldots, x_n). \tag{6.11}$$

We might also like to determine the standard error in the estimate $\hat{\theta}$ and the bias. The bootstrap method can provide an estimate of this when analytical methods fail. The method is also suitable for situations when the estimator $\hat{\theta} = t(\mathrm{x})$ is complicated.

To get estimates of bias or standard error of a statistic, we obtain B bootstrap samples by sampling with replacement from the original sample. For every bootstrap sample, we calculate the same statistic to obtain the ***bootstrap replications*** of $\hat{\theta}$, as follows

$$\hat{\theta}^{*b} = t(\mathrm{x}^{*b}); \qquad b = 1, \ldots, B. \tag{6.12}$$

These B bootstrap replicates provide us with an estimate of the distribution of $\hat{\theta}$. This is similar to what we did in the previous section, except that we are not making any assumptions about the distribution for the original sample. Once we have the bootstrap replicates in Equation 6.12, we can use them to understand the distribution of the estimate.

The steps for the basic bootstrap methodology are given here, with detailed procedures for finding specific characteristics of $\hat{\theta}$ provided later. The issue of how large to make B is addressed with each application of the bootstrap.

PROCEDURE - BASIC BOOTSTRAP

1. Given a random sample, $\mathbf{x} = (x_1, \ldots, x_n)$, calculate $\hat{\theta}$.
2. Sample with replacement from the original sample to get $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$.
3. Calculate the same statistic using the bootstrap sample in step 2 to get, $\hat{\theta}^{*b}$.
4. Repeat steps 2 through 3, B times.
5. Use this estimate of the distribution of $\hat{\theta}$ (i.e., the bootstrap replicates) to obtain the desired characteristic (e.g., standard error, bias or confidence interval).

Efron and Tibshirani [1993] discuss a method called the ***parametric bootstrap***. In this case, the data analyst makes an assumption about the distribution that generated the original sample. Parameters for that distribution are estimated from the sample, and resampling (in step 2) is done using the assumed distribution and the estimated parameters. The parametric bootstrap is closer to the Monte Carlo methods described in the previous section.

For instance, say we have reason to believe that the data come from an exponential distribution with parameter λ. We need to estimate the variance and use

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (6.13)$$

as the estimator. We can use the parametric bootstrap as outlined above to understand the behavior of $\hat{\theta}$. Since we assume an exponential distribution for the data, we estimate the parameter λ from the sample to get $\hat{\lambda}$. We then resample from an exponential distribution with parameter $\hat{\lambda}$ to get the bootstrap samples. The reader is asked to implement the parametric bootstrap in the exercises.

Bootstrap Estimate of Standard Error

When our goal is to estimate the standard error of $\hat{\theta}$ using the bootstrap method, we proceed as outlined in the previous procedure. Once we have the estimated distribution for $\hat{\theta}$, we use it to estimate the standard error for $\hat{\theta}$. This estimate is given by

$$\hat{SE}_B(\hat{\theta}) = \left\{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \overline{\hat{\theta}^{*}})^2\right\}^{\frac{1}{2}}, \qquad (6.14)$$

where

$$\overline{\hat{\theta}^{*}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}. \qquad (6.15)$$

Note that Equation 6.14 is just the sample standard deviation of the bootstrap replicates, and Equation 6.15 is the sample mean of the bootstrap replicates.

Efron and Tibshirani [1993] show that the number of bootstrap replicates B should be between 50 and 200 when estimating the standard error of a statistic. Often the choice of B is dictated by the computational complexity of $\hat{\theta}$, the sample size n, and the computer resources that are available. Even using a small value of B, say $B = 25$, the analyst will gain information about the variability of $\hat{\theta}$. In most cases, taking more than 200 bootstrap replicates to estimate the standard error is unnecessary.

The procedure for finding the bootstrap estimate of the standard error is given here and is illustrated in Example 6.9

PROCEDURE - BOOTSTRAP ESTIMATE OF THE STANDARD ERROR

1. Given a random sample, $\mathbf{x} = (x_1, \ldots, x_n)$, calculate the statistic $\hat{\theta}$.
2. Sample with replacement from the original sample to get $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$.
3. Calculate the same statistic using the sample in step 2 to get the bootstrap replicates, $\hat{\theta}^{*b}$.
4. Repeat steps 2 through 3, B times.
5. Estimate the standard error of $\hat{\theta}$ using Equations 6.14 and 6.15.

Example 6.9

The lengths of the forearm (in inches) of 140 adult males are contained in the file **forearm** [Hand, et al., 1994]. We use these data to estimate the skewness of the population. We then estimate the standard error of this statistic using the bootstrap method. First we load the data and calculate the skewness.

```
load forearm
% Sample with replacement from this.
% First get the sample size.
n = length(forearm);
B = 100;% number of bootstrap replicates
% Get the value of the statistic of interest.
theta = skewness(forearm);
```

The estimated skewness in the **forearm** data is -0.11. To implement the bootstrap, we use the MATLAB Statistics Toolbox function **unidrnd** to sample with replacement from the original sample. The corresponding function from the Computational Statistics Toolbox can also be used. The output from this function will be indices from 1 to n that point to what observations have been selected for the bootstrap sample.

```
% Use unidrnd to get the indices to the resamples.
% Note that each column corresponds to indices
% for a bootstrap resample.
inds = unidrnd(n,n,B);
% Extract these from the data.
xboot = forearm(inds);
% We can get the skewness for each column using the
% MATLAB Statistics Toolbox function skewness.
```

```
thetab = skewness(xboot);
seb = std(thetab);
```

From this we get an estimated standard error in the skewness of 0.14. Efron and Tibshirani [1993] recommend that one look at histograms of the bootstrap replicates as a useful tool for understanding the distribution of $\hat{\theta}$. We show the histogram in Figure 6.6.

The MATLAB Statistics Toolbox has a function called **bootstrp** that returns the bootstrap replicates. We now show how to get the bootstrap estimate of standard error using this function.

```
% Now show how to do it with MATLAB Statistics Toolbox
% function: bootstrp.
Bmat = bootstrp(B,'skewness',forearm);
% What we get back are the bootstrap replicates.
% Get an estimate of the standard error.
sebmat = std(Bmat);
```

Note that one of the arguments to **bootstrp** is a string representing the function that calculates the statistics. From this, we get an estimated standard error of 0.12.

❑

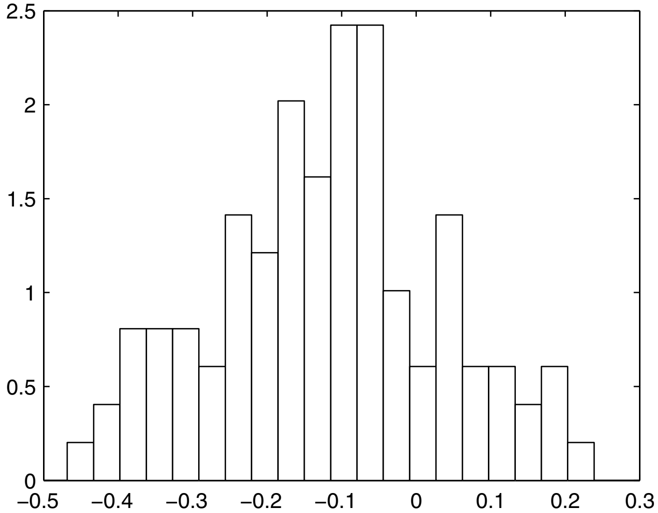

FIGURE 6.6
This is a histogram for the bootstrap replicates in Example 6.9. This shows the estimated distribution of the sample skewness of the **forearm** data.

Bootstrap Estimate of Bias

The standard error of an estimate is one measure of its performance. Bias is another quantity that measures the statistical accuracy of an estimate. From Chapter 3, the bias is defined as the difference between the expected value of the statistic and the parameter,

$$\text{bias}(T) = E[T] - \theta . \tag{6.16}$$

The expectation in Equation 6.16 is taken with respect to the true distribution F. To get the bootstrap estimate of bias, we use the empirical distribution $\hat{F}$ as before. We resample from the empirical distribution and calculate the statistic using each bootstrap resample, yielding the bootstrap replicates $\hat{\theta}^{*b}$. We use these to estimate the bias from the following:

$$\widehat{\text{bias}}_B = \overline{\hat{\theta}^*} - \hat{\theta} , \tag{6.17}$$

where $\overline{\hat{\theta}^*}$ is given by the mean of the bootstrap replicates (Equation 6.15).

Presumably, one is interested in the bias in order to correct for it. The bias-corrected estimator is given by

$$\overset{\frown}{\theta} = \hat{\theta} - \widehat{\text{bias}}_B . \tag{6.18}$$

Using Equation 6.17 in Equation 6.18, we have

$$\overset{\frown}{\theta} = 2\hat{\theta} - \overline{\hat{\theta}^*} . \tag{6.19}$$

More bootstrap samples are needed to estimate the bias, than are required to estimate the standard error. Efron and Tibshirani [1993] recommend that $B \geq 400$.

It is useful to have an estimate of the bias for $\hat{\theta}$, but caution should be used when correcting for the bias. Equation 6.19 will hopefully yield a less biased estimate, but it could turn out that $\overset{\frown}{\theta}$ will have a larger variation or standard error. It is recommended that if the estimated bias is small relative to the estimate of standard error (both of which can be estimated using the bootstrap method), then the analyst should not correct for the bias [Efron and Tibshirani, 1993]. However, if this is not the case, then perhaps some other, less biased, estimator should be used to estimate the parameter θ.

PROCEDURE - BOOTSTRAP ESTIMATE OF THE BIAS

1. Given a random sample, $\mathbf{x} = (x_1, \ldots, x_n)$, calculate the statistic $\hat{\theta}$.

2. Sample with replacement from the original sample to get $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$.
3. Calculate the same statistic using the sample in step 2 to get the bootstrap replicates, $\hat{\theta}^{*b}$.
4. Repeat steps 2 through 3, B times.
5. Using the bootstrap replicates, calculate $\overline{\hat{\theta}^*}$.
6. Estimate the bias of $\hat{\theta}$ using Equation 6.17.

Example 6.10

We return to the **forearm** data of Example 6.9, where now we want to estimate the bias in the sample skewness. We use the same bootstrap replicates as before, so all we have to do is to calculate the bias using Equation 6.17.

```
% Use the same replicates from before.
% Evaluate the mean using Equation 6.15.
meanb = mean(thetab);
% Now estimate the bias using Equation 6.17.
biasb = meanb - theta;
```

We have an estimated bias of -0.011. Note that this is small relative to the standard error.

❑

In the next chapter, we discuss another method for estimating the bias and the standard error of a statistic called the jackknife. The jackknife method is related to the bootstrap. However, since it is based on the reuse or partitioning of the original sample rather than resampling, we do not include it here.

Bootstrap Confidence Intervals

There are several ways of constructing confidence intervals using the bootstrap. We discuss three of them here: the standard interval, the bootstrap-t interval and the percentile method. Because it uses the jackknife procedure, an improved bootstrap confidence interval called the BC_a will be presented in the next chapter.

Bootstrap Standard Confidence Interval

The ***bootstrap standard confidence interval*** is based on the parametric form of the confidence interval that was discussed in Section 6.2. We showed that the $(1-\alpha)\cdot 100\%$ confidence interval for the mean can be found using

$$P\left(\bar{X} - z^{(1-\alpha/2)}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - z^{(\alpha/2)}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha. \qquad (6.20)$$

Similar to this, the bootstrap standard confidence interval is given by

$$(\hat{\theta} - z^{(1-\alpha/2)} SE_{\hat{\theta}}, \hat{\theta} - z^{(\alpha/2)} SE_{\hat{\theta}}) , \tag{6.21}$$

where $SE_{\hat{\theta}}$ is the standard error for the statistic $\hat{\theta}$ obtained using the bootstrap [Mooney and Duval, 1993]. The confidence interval in Equation 6.21 can be used when the distribution for $\hat{\theta}$ is normally distributed or the normality assumption is plausible. This is easily coded in MATLAB using previous results and is left as an exercise for the reader.

Bootstrap-t Confidence Interval

The second type of confidence interval using the bootstrap is called the ***bootstrap-t***. We first generate B bootstrap samples, and for each bootstrap sample the following quantity is computed:

$$z^{*b} = \frac{\hat{\theta}^{*b} - \hat{\theta}}{\hat{SE}^{*b}} . \tag{6.22}$$

As before, $\hat{\theta}^{*b}$ is the bootstrap replicate of $\hat{\theta}$, but $\hat{SE}^{*b}$ is the estimated standard error of $\hat{\theta}^{*b}$ for that bootstrap sample. If a formula exists for the standard error of $\hat{\theta}^{*b}$, then we can use that to determine the denominator of Equation 6.22. For instance, if $\hat{\theta}$ is the mean, then we can calculate the standard error as explained in Chapter 3. However, in most situations where we have to resort to using the bootstrap, these formulas are not available. One option is to use the bootstrap method of finding the standard error, keeping in mind that you are estimating the standard error of $\hat{\theta}^{*b}$ using the bootstrap sample $\mathbf{x}^{*b}$. In other words, one resamples with replacement from the bootstrap sample $\mathbf{x}^{*b}$ to get an estimate of $\hat{SE}^{*b}$.

Once we have the B bootstrapped z^{*b} values from Equation 6.22, the next step is to estimate the quantiles needed for the endpoints of the interval. The $\alpha/2$-th quantile, denoted by $\hat{t}^{(\alpha/2)}$ of the z^{*b}, is estimated by

$$\alpha/2 = \frac{\#(z^{*b} \le \hat{t}^{(\alpha/2)})}{B} . \tag{6.23}$$

This says that the estimated quantile is the $\hat{t}^{(\alpha/2)}$ such that $100 \cdot \alpha/2$ % of the points z^{*b} are less than this number. For example, if $B = 100$ and $\alpha/2 = 0.05$, then $\hat{t}^{(0.05)}$ could be estimated as the fifth largest value of the z^{*b} $(B \cdot \alpha/2 = 100 \cdot 0.05 = 5)$. One could also use the quantile estimates discussed previously in Chapter 3 or some other suitable estimate.

We are now ready to calculate the bootstrap-*t* confidence interval. This is given by

$$(\hat{\theta} - \hat{t}^{(1-\alpha/2)} \cdot \hat{SE}_{\hat{\theta}}, \hat{\theta} - \hat{t}^{(\alpha/2)} \cdot \hat{SE}_{\hat{\theta}}), \qquad (6.24)$$

where $\hat{SE}$ is an estimate of the standard error of $\hat{\theta}$. The bootstrap-t interval is suitable for location statistics such as the mean or quantiles. However, its accuracy for more general situations is questionable [Efron and Tibshirani, 1993]. The next method based on the bootstrap percentiles is more reliable.

PROCEDURE - BOOTSTRAP-T CONFIDENCE INTERVAL

1. Given a random sample, $\mathbf{x} = (x_1, \ldots, x_n)$, calculate $\hat{\theta}$.
2. Sample with replacement from the original sample to get $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$.
3. Calculate the same statistic using the sample in step 2 to get $\hat{\theta}^{*b}$.
4. Use the bootstrap sample $\mathbf{x}^{*b}$ to get the standard error of $\hat{\theta}^{*b}$. This can be calculated using a formula or estimated by the bootstrap.
5. Calculate z^{*b} using the information found in steps 3 and 4.
6. Repeat steps 2 through 5, B times, where $B \geq 1000$.
7. Order the z^{*b} from smallest to largest. Find the quantiles $\hat{t}^{(1-\alpha/2)}$ and $\hat{t}^{(\alpha/2)}$.
8. Estimate the standard error $\hat{SE}_{\hat{\theta}}$ of $\hat{\theta}$ using the B bootstrap replicates of $\hat{\theta}^{*b}$ (from step 3).
9. Use Equation 6.24 to get the confidence interval.

The number of bootstrap replicates that are needed is quite large for confidence intervals. It is recommended that B should be 1000 or more. If no formula exists for calculating the standard error of $\hat{\theta}^{*b}$, then the bootstrap method can be used. This means that there are two levels of bootstrapping: one for finding the $\hat{SE}^{*b}$ and one for finding the z^{*b}, which can greatly increase the computational burden. For example, say that $B = 1000$ and we use 50 bootstrap replicates to find $\hat{SE}^{*b}$, then this results in a total of 50,000 resamples.

Example 6.11

Say we are interested in estimating the variance of the **`forearm`** data, and we decide to use the following statistic,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which is the sample second central moment. We write our own simple function called **mom** (included in the Computational Statistics Toolbox) to estimate this.

```
% This function will calculate the sample 2nd
% central moment for a given sample vector x.
function mr = mom(x)
n = length(x);
mu = mean(x);
mr =  (1/n)*sum((x-mu).^2);
```

We use this function as an input argument to **bootstrp** to get the bootstrap-*t* confidence interval. The MATLAB code given below also shows how to get the bootstrap estimate of standard error for each bootstrap sample. First we load the data and get the observed value of the statistic.

```
load forearm
n = length(forearm);
alpha = 0.1;
B = 1000;
thetahat = mom(forearm);
```

Now we get the bootstrap replicates using the function **bootstrp**. One of the optional output arguments from this function is a matrix of indices for the resamples. As shown below, each column of the output **bootsam** contains the indices to a bootstrap sample. We loop through all of the bootstrap samples to estimate the standard error of the bootstrap replicate using that resample.

```
% Get the bootstrap replicates and samples.
[bootreps, bootsam] = bootstrp(B,'mom',forearm);
% Set up some storage space for the SE’s.
sehats = zeros(size(bootreps));
% Each column of bootsam contains indices
% to a bootstrap sample.
for i = 1:B
    % Extract the sample from the data.
     xstar = forearm(bootsam(:,i));
   bvals(i) = mom(xstar);
   % Do bootstrap using that sample to estimate SE.
   sehats(i) = std(bootstrp(25,'mom',xstar));
end
zvals = (bootreps - thetahat)./sehats;
```

Then we get the estimate of the standard error that we need for the endpoints of the interval.

```
% Estimate the SE using the bootstrap.
SE = std(bootreps);
```

Now we get the quantiles that we need for the interval given in Equation 6.24 and calculate the interval.

```
% Get the quantiles.
k = B*alpha/2;
szval = sort(zvals);
tlo = szval(k);
thi = szval(B-k);
% Get the endpoints of the interval.
blo = thetahat - thi*SE;
bhi = thetahat - tlo*SE;
```

The bootstrap-*t* interval for the variance of the `forearm` data is $(1.00, 1.57)$.
❑

Bootstrap Percentile Interval

An improved bootstrap confidence interval is based on the quantiles of the distribution of the bootstrap replicates. This technique has the benefit of being more stable than the bootstrap-*t*, and it also enjoys better theoretical coverage properties [Efron and Tibshirani, 1993]. The ***bootstrap percentile confidence interval*** is

$$(\hat{\theta}_B^{*(\alpha/2)}, \hat{\theta}_B^{*(1-\alpha/2)}), \tag{6.25}$$

where $\hat{\theta}_B^{*(\alpha/2)}$ is the $\alpha/2$ quantile in the bootstrap distribution of $\hat{\theta}^*$. For example, if $\alpha/2 = 0.025$ and $B = 1000$, then $\hat{\theta}_B^{*(0.025)}$ is the $\hat{\theta}^{*b}$ in the 25th position of the ordered bootstrap replicates. Similarly, $\hat{\theta}_B^{*(0.975)}$ is the replicate in position 975. As discussed previously, some other suitable estimate for the quantile can be used.

The procedure is the same as the general bootstrap method, making it easy to understand and to implement. We outline the steps below.

PROCEDURE - BOOTSTRAP PERCENTILE INTERVAL

1. Given a random sample, $\mathbf{x} = (x_1, \ldots, x_n)$, calculate $\hat{\theta}$.
2. Sample with replacement from the original sample to get $\mathbf{x}^{*b} = (x_1^{*b}, \ldots, x_n^{*b})$.
3. Calculate the same statistic using the sample in step 2 to get the bootstrap replicates, $\hat{\theta}^{*b}$.
4. Repeat steps 2 through 3, B times, where $B \geq 1000$.
5. Order the $\hat{\theta}^{*b}$ from smallest to largest.
6. Calculate $B \cdot \alpha/2$ and $B \cdot (1 - \alpha/2)$.

7. The lower endpoint of the interval is given by the bootstrap replicate that is in the $B \cdot \alpha/2$-th position of the ordered $\hat{\theta}^{*b}$, and the upper endpoint is given by the bootstrap replicate that is in the $B \cdot (1 - \alpha/2)$-th position of the same ordered list. Alternatively, using quantile notation, the lower endpoint is the estimated quantile $\hat{q}_{\alpha/2}$ and the upper endpoint is the estimated quantile $\hat{q}_{1-\alpha/2}$, where the estimates are taken from the bootstrap replicates.

Example 6.12

Let's find the bootstrap percentile interval for the same `forearm` data. The confidence interval is easily found from the bootstrap replicates, as shown below.

```
% Use Statistics Toolbox function
% to get the bootstrap replicates.
bvals = bootstrp(B,'mom',forearm);
% Find the upper and lower endpoints
k = B*alpha/2;
sbval = sort(bvals);
blo = sbval(k);
bhi = sbval(B-k);
```

This interval is given by $(1.03, 1.45)$, which is slightly narrower than the bootstrap-*t* interval from Example 6.11.

❑

So far, we discussed three types of bootstrap confidence intervals. The standard interval is the easiest and assumes that $\hat{\theta}$ is normally distributed. The bootstrap-*t* interval estimates the standardized version of $\hat{\theta}$ from the data, avoiding the normality assumptions used in the standard interval. The percentile interval is simple to calculate and obtains the endpoints directly from the bootstrap estimate of the distribution for $\hat{\theta}$. It has another advantage in that it is range-preserving. This means that if the parameter θ can take on values in a certain range, then the confidence interval will reflect that. This is not always the case with the other intervals.

According to Efron and Tibshirani [1993], the bootstrap-*t* interval has good coverage probabilities, but does not perform well in practice. The bootstrap percentile interval is more dependable in most situations, but does not enjoy the good coverage property of the bootstrap-*t* interval. There is another bootstrap confidence interval, called the BC_a interval, that has both good coverage and is dependable. This interval is described in the next chapter.

The bootstrap estimates of bias and standard error are also random variables, and they have their own error associated with them. So, how accurate are they? In the next chapter, we discuss how one can use the jackknife method to evaluate the error in the bootstrap estimates.

As with any method, the bootstrap is not appropriate in every situation. When analytical methods are available to understand the uncertainty associated with an estimate, then those are more efficient than the bootstrap. In what situations should the analyst use caution in applying the bootstrap? One important assumption that underlies the theory of the bootstrap is the notion that the empirical distribution function is representative of the true population distribution. If this is not the case, then the bootstrap will not yield reliable results. For example, this can happen when the sample size is small or the sample was not gathered using appropriate random sampling techniques. Chernick [1999] describes other examples from the literature where the bootstrap should not be used. We also address a situation in Chapter 7 where the bootstrap fails. This can happen when the statistic is nonsmooth, such as the median.

6.5 MATLAB Code

We include several functions with the Computational Statistics Toolbox that implement some of the bootstrap techniques discussed in this chapter. These are listed in Table 6.2. Like **bootstrp**, these functions have an input argument that specifies a MATLAB function that calculates the statistic.

TABLE 6.2

List of MATLAB Functions for Chapter 6

Purpose	MATLAB Function
General bootstrap: resampling, estimates of standard error and bias	**csboot** **bootstrp**
Constructing bootstrap confidence Intervals	**csbootint** **csbooperint** **csbootbca**

As we saw in the examples, the MATLAB Statistics Toolbox has a function called **bootstrp** that will return the bootstrap replicates from the input argument **bootfun** (e.g., **mean**, **std**, **var**, etc.). It takes an input data set, finds the bootstrap resamples, applies the **bootfun** to the resamples, and stores the replicate in the first row of the output argument. The user can get two outputs from the function: the bootstrap replicates and the indices that correspond to the points selected in the resample.

There is a Bootstrap MATLAB Toolbox written by Zoubir and Iskander at the Curtin University of Technology. It is available for download at

www.atri.curtin.edu.au/csp. It requires the MATLAB Statistics Toolbox and has a postscript version of the reference manual.

Other software exists for Monte Carlo simulation as applied to statistics. The Efron and Tibshirani [1993] book has a description of S code for implementing the bootstrap. This code, written by the authors, can be downloaded from the statistics archive at Carnegie-Mellon University that was mentioned in Chapter 1. Another software package that has some of these capabilities is called Resampling Stats® [Simon, 1999], and information on this can be found at **www.resample.com**. Routines are available from Resampling Stats for MATLAB [Kaplan, 1999] and Excel.

6.6 Further Reading

Mooney [1997] describes Monte Carlo simulation for inferential statistics that is written in a way that is accessible to most data analysts. It has some excellent examples of using Monte Carlo simulation for hypothesis testing using multiple experiments, assessing the behavior of an estimator, and exploring the distribution of a statistic using graphical techniques. The text by Gentle [1998] has a chapter on performing Monte Carlo studies in statistics. He discusses how simulation can be considered as a scientific experiment and should be held to the same high standards. Hoaglin and Andrews [1975] provide guidelines and standards for reporting the results from computations. Efron and Tibshirani [1991] explain several computational techniques, written at a level accessible to most readers. Other articles describing Monte Carlo inferential methods can be found in Joeckel [1991], Hope [1968], Besag and Diggle [1977], Diggle and Gratton [1984], Efron [1979], Efron and Gong [1983], and Teichroew [1965].

There has been a lot of work in the literature on bootstrap methods. Perhaps the most comprehensive and easy to understand treatment of the topic can be found in Efron and Tibshirani [1993]. Efron's [1982] earlier monogram on resampling techniques describes the jackknife, the bootstrap and cross-validation. A more recent book by Chernick [1999] gives an updated description of results in this area, and it also has an extensive bibliography (over 1,600 references!) on the bootstrap. Hall [1992] describes the connection between Edgeworth expansions and the bootstrap. A volume of papers on the bootstrap was edited by LePage and Billard [1992], where many applications of the bootstrap are explored. Politis, Romano, andWolf [1999] present subsampling as an alternative to the bootstrap. A subset of articles that present the theoretical justification for the bootstrap are Efron [1981, 1985, 1987]. The paper by Boos and Zhang [2000] looks at a way to ease the computational burden of Monte Carlo estimation of the power of tests that uses resampling methods. For a nice discussion on the coverage of the bootstrap percentile confidence interval, see Polansky [1999].

Exercises

6.1. Repeat Example 6.1 where the population standard deviation for the travel times to work is $\sigma_X = 5$ minutes. Is $\bar{x} = 47.2$ minutes still consistent with the null hypothesis?

6.2. Using the information in Example 6.3, plot the probability of Type II error as a function of μ. How does this compare with Figure 6.2?

6.3. Would you reject the null hypothesis in Example 6.4 if $\alpha = 0.10$?

6.4. Using the same value for the sample mean, repeat Example 6.3 for different sample sizes of $n = 50, 100, 200$. What happens to the curve showing the power as a function of the true mean as the sample size changes?

6.5. Repeat Example 6.6 using a two-tail test. In other words, test for the alternative hypothesis that the mean is not equal to 454.

6.6. Repeat Example 6.8 for larger *M*. Does the estimated Type I error get closer to the true value?

6.7. Write MATLAB code that implements the parametric bootstrap. Test it using the **forearm** data. Assume that the normal distribution is a reasonable model for the data. Use your code to get a bootstrap estimate of the standard error and the bias of the coefficient of skewness and the coefficient of kurtosis. Get a bootstrap percentile interval for the sample central second moment using your parametric bootstrap approach.

6.8. Write MATLAB code that will get the bootstrap standard confidence interval. Use it with the **forearm** data to get a confidence interval for the sample central second moment. Compare this interval with the ones obtained in the examples and in the previous problem.

6.9. Use your program from problem 6.8 and the **forearm** data to get a bootstrap confidence interval for the mean. Compare this to the theoretical one.

6.10. The **remiss** data set contains the remission times for 42 leukemia patients. Some of the patients were treated with the drug called 6-mercaptopurine (**mp**), and the rest were part of the control group (**control**). Use the techniques from Chapter 5 to help determine a suitable model (e.g., Weibull, exponential, etc.) for each group. Devise a Monte Carlo hypothesis test to test for the equality of means between the two groups [Hand, et al., 1994; Gehan, 1965]. Use the *p*-value approach.

6.11. Load the **lawpop** data set [Efron and Tibshirani, 1993]. These data contain the average scores on the LSAT (**lsat**) and the corresponding average undergraduate grade point average (**gpa**) for the 1973 freshman class at 82 law schools. Note that these data constitute the entire population. The data contained in **law** comprise a random sample of 15 of these classes. Obtain the true population variances for the **lsat** and the **gpa**. Use the sample in **law** to estimate the population variance using the sample central second moment. Get bootstrap estimates of the standard error and the bias in your estimate of the variance. Make some comparisons between the known population variance and the estimated variance.

6.12. Using the **lawpop** data, devise a test statistic to test for the significance of the correlation between the LSAT scores and the corresponding grade point averages. Get a random sample from the population, and use that sample to test your hypothesis. Do a Monte Carlo simulation of the Type I and Type II error of the test you devise.

6.13. In 1961, 16 states owned the retail liquor stores. In 26 others, the stores were owned by private citizens. The data contained in **whisky** reflect the price (in dollars) of a fifth of whisky from these 42 states. Note that this represents the population, not a sample. Use the **whisky** data to get an appropriate bootstrap confidence interval for the median price of whisky at the state owned stores and the median price of whisky at the privately owned stores. First get the random sample from each of the populations, and then use the bootstrap with that sample to get the confidence intervals. Do a Monte Carlo study where you compare the confidence intervals for different sample sizes. Compare the intervals with the known population medians [Hand, et al., 1994].

6.14. The **quakes** data [Hand, et al., 1994] give the time in days between successive earthquakes. Use the bootstrap to get an appropriate confidence interval for the average time between earthquakes.